基于阅读技巧识别和双通道融合机制的机器阅读理解方法

彭 伟[1,2] 胡 玥[1,3] 李运鹏[1,3] 谢玉强[1,3] 牛晨旭[1,3]

摘 要 机器阅读理解任务旨在要求系统对给定文章进行理解, 然后对给定问题进行回答. 先前的工作重点聚焦在问题和文章间的交互信息, 忽略了对问题进行更加细粒度的分析 (如问题所考察的阅读技巧是什么?). 受先前研究的启发, 人类对于问题的理解是一个多维度的过程. 首先, 人类需要理解问题的上下文信息; 然后, 针对不同类型问题, 识别其需要使用的阅读技巧; 最后, 通过与文章交互回答出问题答案. 针对这些问题, 提出一种基于阅读技巧识别和双通道融合的机器阅读理解方法, 对问题进行更加细致的分析, 从而提高模型回答问题的准确性. 阅读技巧识别器通过对比学习的方法, 能够显式地捕获阅读技巧的语义信息. 双通道融合机制将问题与文章的交互信息和阅读技巧的语义信息进行深层次的融合, 从而达到辅助系统理解问题和文章的目的. 为了验证该模型的效果, 在 FairytaleQA 数据集上进行实验, 实验结果表明, 该方法实现了在机器阅读理解任务和阅读技巧识别任务上的最好效果.

关键词 机器阅读理解, 阅读技巧识别, 对比学习, 双通道融合机制

引用格式 彭伟, 胡玥, 李运鹏, 谢玉强, 牛晨旭. 基于阅读技巧识别和双通道融合机制的机器阅读理解方法. 自动化学报, 2024, **50**(5): 958−969

DOI 10.16383/j.aas.c220983

A Machine Reading Comprehension Approach Based on Reading Skill Recognition and Dual Channel Fusion Mechanism

PENG Wei[1,2] HU Yue[1,3] LI Yun-Peng[1,3] XIE Yu-Qiang[1,3] NIU Chen-Xu[1,3]

Abstract Machine reading comprehension task aims to require the system to understand a given passage and then answer a question. Previous researches focus on the interaction between questions and passages. However, they neglect to make a more granular analysis of the questions, e.g., what is the reading skill examined by the questions? Inspired by the previous reading comprehension literature, the understanding of questions is a multi-dimensional process where humans first need to understand the context semantics of the question, then identify the reading skills they need to use for different types of questions, and finally answer the question. In the end, we propose a machine reading comprehension method based on reading skill recognition and dual channel fusion mechanism to make a comprehensive analysis of questions, so as to improve the accuracy of the model in answering questions. Specifically, the reading skill recognizer can capture the semantic representations of reading skills through contrastive learning. The dual channel fusion mechanism deeply integrates the contextual information and the semantic representations of reading skills, so as to help the system understand the question and passage. To verify the effectivenesss of the model, we conduct experiments on the FairytaleQA dataset. The experimental results show that the proposed method achieves the state-of-the-art performance on machine reading comprehension task and reading skill recognition task.

Key words Machine reading comprehension, reading skill recognition, contrastive learning, dual channel fusion machanism

Citation Peng Wei, Hu Yue, Li Yun-Peng, Xie Yu-Qiang, Niu Chen-Xu. A machine reading comprehension approach based on reading skill recognition and dual channel fusion mechanism. *Acta Automatica Sinica*, 2024, **50**(5): 958−969

随着深度学习的发展, 机器阅读理解[1−3] 已经取得了极大进步[4−6], 其是一项测试机器理解自然语言的任务, 要求模型在给定的一篇文本内容的基础上, 对相应的问题作出回答. 不同于传统的检索式问答系统, 机器阅读理解更加考查模型对语言的理

收稿日期 2022-12-20 录用日期 2023-07-22
Manuscript received December 20, 2022; accepted July 22, 2023
国家自然科学基金 (62006222, U21B2009), 中国科学院战略性先导研究计划 (XDC02030400) 资助
Supported by National Natural Science Foundation of China (62006222, U21B2009) and Strategic Priority Research Program of Chinese Academy of Science (XDC02030400)
本文责任编委 金连文
Recommended by Associate Editor JIN Lian-Wen
1. 中国科学院信息工程研究所 北京 100085 2. 中关村实验室 北京 100080 3. 中国科学院大学网络空间安全学院 北京 100085
1. Institute of Information Engineering, Chinese Academy of Sciences, Beijing 100085 2. Zhongguancun Laboratory, Beijing 100080 3. School of Cyber Security, University of Chinese Academy of Sciences, Beijing 100085

解和推理能力[7]. 此外, 机器阅读理解系统有广泛的应用前景, 包括语音助手、智能客户服务系统和搜索引擎等.

近年来, 许多研究工作聚焦在建模问题和文章间的交互过程, 这些交互过程可分为浅层次交互[2]和深层次交互[8] 2 种. 如 Seo 等[2] 首次引入双向注意力流, 以建模问题对文章、文章对问题的语义感知能力; Peng 等[4] 从多步推理角度, 挖掘文章中与问题相关的关键信息进行问答; Liao 等[9] 提出利用异质图的方法, 使模型进行深层次交互, 充分地挖掘问题与文章间的依赖关系. 这些方法多是通过问题和文章间的交互信息和关键证据进行问答, 很少关注问题中更加细粒度的信息, 如阅读技巧识别等.

受文献 [10−12] 启发, 人类对于问题的理解是一个多维度过程[11]. 在这个过程中, 需要首先对问题的上下文信息进行理解[10], 然后针对不同问题、利用不同阅读技巧进行回答[13]. Purves 等[14] 将阅读任务分解为 3 个阶段, 每个阶段检测不同的阅读技巧, 这些阅读技巧可以以一种确定性方式来回答对应的问题[15]. 如当问题考察的是时间和地点相关内容, 那么机器就应该更多地关注文章中的时间和地点实体. 通过在 FairytaleQA 数据集中的一个例子描述上述过程 (见图 1). 给定一个问题和一篇文章, 系统需要在步骤 1 根据问题来分析其背后所涉及的阅读技巧; 在步骤 2, 根据挖掘出的阅读技巧和文章信息输出答案. 在图 1 的例子中, 阅读技巧被识别成“动作”, 表示这个技巧能被用来回答人物的行为或与行为相关的信息. 所以, 机器会更加关注到人物所做的事件, 以此来生成答案“他就会被赶走, 被射杀, ……”. 总之, 对问题的理解不应该仅局限其上下文信息上, 而且需要去挖掘隐藏在问题背后的阅读技巧进行识别.

为解决上述问题, 本文提出基于阅读技巧识别和双通道融合的机器阅读理解方法, 对问题进行更细致地分析. 本文模型包含上下文编码器、阅读技巧识别器、双通道融合机制和答案生成器. 其中, 上下文编码器的作用是建模问题与文章上下文信息; 阅读技巧识别器通过对比学习方法, 捕获阅读技巧的语义表示; 双通道融合机制将问题与文章上下文信息和阅读技巧的语义表示进行深层次地融合; 答案生成器根据融合后的表示, 生成准确答案. 本文的主要贡献有以下 4 点:

1) 模拟人类阅读过程. 本文基于阅读技巧识别和双通道融合的机器阅读理解方法能加强模型对问题的细致分析, 提高模型回答问题的准确性.

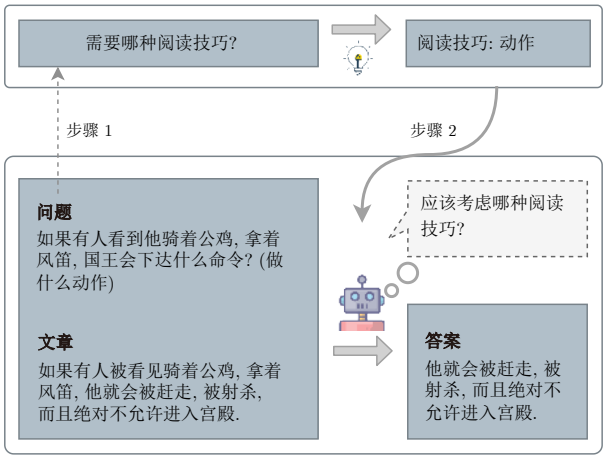


图 1　在 FairytaleQA 数据集中的一个例子

Fig. 1　An example in FairytaleQA dataset

2) 为提高阅读技巧识别的精确度, 本文提出一种有监督对比学习的方法, 准确地捕获阅读技巧的语义表示.

3) 为更好地融入问题和文章的上下文信息以及阅读技巧的语义表示, 本文提出 3 种不同的双通道融合机制.

4) 实验结果表明, 本文方法在机器阅读理解任务和阅读技巧识别任务上实现了最好效果.

1　研究现状

1.1　机器阅读理解数据集和模型

近年来, 随着大规模数据集的发布, 机器阅读理解受到了广泛关注. 数据集可分为完形填空、多项选择、片段抽取和自由问答 4 种形式. 如 CNN & Daily Mail[16]、SQuAD[17]、RACE[18]、NarrativeQA[19] 和 FairytaleQA[6]. 完形填空数据集 (如 CNN & Daily Mail) 要求系统用正确单词填空. 多项选择数据集 (如 RACE) 要求模型从多个选项中选择一个答案, 然后将任务转到抽取式机器阅读理解任务 (如 SQuAD) 中, 抽取的答案来自给定上下文的文本. 最近, 自由问答形式的机器阅读理解任务受到了广泛的关注, 其是一种最复杂的机器阅读理解任务, 因为该任务的答案形式没有任何限制, 可以从文中抽取, 也可以进行归纳总结得出最终的答案. FairytaleQA 数据集就属于这类任务, 其中包含丰富的数据标注信息, 如阅读技巧标签、答案是否来源于原始文本等. 本文将重点关注 FairytaleQA 数据集中的问答任务.

机器阅读理解模型包含嵌入层、特征抽取层、文章问题交互层和答案预测层 4 个主要部分[7]. 主流方法聚焦在嵌入层[20] 和文章问题交互层[2, 9]. 如 Zhang 等[21] 在机器阅读理解任务中, 引入基于语法树信息, 通过信息增强方式辅助问答任务; Peng 等[4] 提出基于证据驱动的推理网络来挖掘问题和文章间的交互信息, 提高阅读理解模型的表现能力; Liao 等[9] 提出利用异质图的方式, 进行深层次的交互. 这些方法多是基于外部信息或通过建模问题与文章间深层次的交互, 来实现机器阅读理解系统. 另外, 部分工作考虑分析问题的类型, 来进一步加强机器阅读理解模型的性能. 如 Kao 等[22] 通过将问题类型及其类型定义结合到输入中, 来改进 FairytaleQA 数据集, 微调 BART (Bidirectional and auto-regressive transformers) 模型, 提高阅读理解系统的效果. Lu 等[23] 提出一种名为考虑问题类型的新方法, 利用模板通过事实三元组来理解自然语言问题, 这些模板可以充分表达用户的意图, 并根据这些意图生成结构化查询辅助问答. 考虑到先前工作忽略了在学习中使用问题类型等问题, 文献 [24] 通过将问题类型与多模态联合表示直接连接, 达到缩小候选答案空间的目的. 本文重点关注显式地捕获阅读技巧的语义信息, 并将其和上下文信息进行深层次融合.

1.2 对比学习

在计算机视觉领域, 基于深度学习的自监督表示学习方法 (如对比学习) 有着广泛的应用[25−27]. 与交叉熵损失相比, 对比学习能够将正样本映射到更紧密的空间, 同时使负样本尽量远离. 与对比学习密切关联的是基于距离度量学习[28] 和三元组损失[29], 其能学习到丰富的句子语义表示. 除了在计算机视觉中的应用, 对比学习也被用于自然语言处理 (Natural language processing, NLP) 任务中. Gao 等[30] 设计一种基于无监督的简单方法, 该方法利用丢弃噪声预测输入句子本身, 并在自然语言推断数据集上设计了一个有监督的方法学习句子的通用表示; Giorgi 等[31] 提出不需要额外标注训练数据的通用句子嵌入的自我监督目标. 此外, 相关研究提出一种利用标签信息的基于有监督的对比学习损失函数, 如文献 [32] 为对比学习损失函数引入了一种新的扩展, 该函数允许每个锚点有多个正样本, 从而使对比学习适应完全监督的设置; Li 等[33] 设计有监督的对比学习模型, 该模型将样本对作为输入, 并使用交叉注意力模块来学习自然语言推理任务中嵌入的句子表示. 本文方法在结构上类似于有监督对比学习中使用的方法, 对监督分类进行了修改.

2 模型介绍

本文模型设计的总体思路为, 首先模型对阅读技巧的表示进行学习, 然后进行机器阅读理解任务. 因此, 本文把整个模型框架分为 2 个训练阶段, 第 1 阶段主要学习阅读技巧的特征表示, 第 2 阶段通过所学习的阅读技巧进行阅读理解任务. 第 1 阶段的核心模块是阅读技巧识别器, 通过有监督对比学习方法准确地捕获阅读技巧的语义表示; 第 2 阶段中的双通道融合机制将上下文信息和技巧语义表示进行融合, 辅助问答.

2.1 问题定义

给定问题 $X^q = \{x_1^q, x_2^q, \cdots, x_N^q\}$ 和文章 $X^p = \{x_1^p, x_2^p, \cdots, x_M^p\}$, 其中 N 和 M 分别表示问题的长度和文章的长度. 定义 X^a 为答案, 对应的阅读技巧为 y^s, 阅读技巧的定义详见第 3.2 节. 本文模型基于上下文信息和技巧语义表示, 输出答案序列 $Y = \{y_1, y_2, \cdots, y_Z\}$.

2.2 本文模型总体结构图

本文模型的总体结构如图 2 所示, 包含上下文编码器、阅读技巧识别器、双通道融合机制和答案生成器 4 个部分. 模型训练分为 2 个阶段. 在第 1 阶段, 对阅读技巧识别器进行预训练, 通过有监督对比学习来学习阅读技巧的语义表示, 这将在第 2 阶段能够对模型回答问题给予指导; 在第 2 阶段, 上下文编码器获得问题和段落间的上下文信息, 然后双通道融合机制将上下文编码器和阅读技巧识别器的输出表示进行深层次融合; 最后, 答案生成器通过自回归的解码方式输出答案.

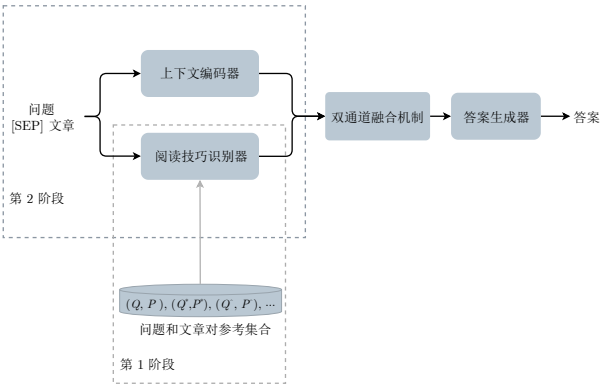


图 2 本文模型总体结构

Fig. 2 Overall structure of our model

2.3　阅读技巧识别器

在第 1 阶段, 通过有监督对比学习对阅读技巧识别器进行预训练. 阅读技巧识别器的目的在于准确捕捉阅读技巧的语义表示, 这些技巧能够用于指导回答不同的问题. 下面将分别介绍正/负样本构造和基于有监督对比学习损失的技巧学习.

2.3.1　正/负样本构造

考虑到 FairytaleQA 数据集中包含问题的丰富标注, 根据给定的样本标签直接构造具有相同技巧标签的样本, 作为正样本; 随机抽取数据集中的其他类别, 作为负样本.

2.3.2　基于有监督对比学习损失的技巧学习

对比学习的目的是将属于同一类样本的语义表示拉近, 并使得不相关的样本尽量远离. 受文献 [33] 启发, 将自监督批量对比学习方法扩展到有监督设置, 以有效利用标签信息. 因此, 本文采用有监督对比学习目标, 准确捕获阅读技巧的语义表示来辅助问答过程. 阅读技巧识别器结构如图 3 所示, 其中左图输入的是正样本, 中间图输入的是原始样本, 右图输入的是负样本.

给定参考集合包含 $\mathcal{B}$ 个实例, 每个实例可表示为 $(X^q, X^p, y^s)_{i\in\mathcal{B}}$, 其中 $i=\{1,\cdots,K\}$ 表示实例索引, K 是批量数据大小. 遵循 BERT (Bidirectional encoder representations from transformers)[34] 构造输入方式, 本文将编码了 2 个句子间的蕴涵信息 [CLS] 作为开始标志符, 并用[SEP] 将问题 X^q 和文章 X^p 进行分隔, 因此输入的总长度 $T=N+M+2$. BERT 上下文编码器将 X^q 和 X^p 作为输入来计算上下文信息, 因此可以得到一系列隐层向量 $(\boldsymbol{h}_1, \boldsymbol{h}_2, \cdots, \boldsymbol{h}_T)$:

$$\boldsymbol{h}_t = \text{BERT}([\text{CLS}], X^q, [\text{SEP}], X^p) \tag{1}$$

式中, $\boldsymbol{h}_t \in \mathbf{R}^d$ 是输入的第 t 个词, d 是 BERT 的隐藏层神经元的维度.

本文将输入的整体语义向量表示定义为 $\boldsymbol{g}$, 可以通过 2 种方式进行计算. 第 1 种简单方式是直接用开始标志符 [CLS] 表示作为 $\boldsymbol{g}$. 第 2 种方式通过对所有的输入特征进行池化操作. 本文使用平均池化操作来捕获句子对 (即问题和文章) 间的相关性:

$$\boldsymbol{g} = \text{Mean-pooling}(\boldsymbol{h}_1, \boldsymbol{h}_2, \cdots, \boldsymbol{h}_T) \tag{2}$$

式中, Mean-pooling 表示平均池化操作. 计算得到该表示后, 通过对比学习方式学习这些样本对间的依赖关系. 在第 1 阶段, 本文随机采用 1 个批处理数据样本 $\mathcal{I}$, 该批处理数据样本包含 K 个实例 $(X^q, X^p, y^s)_{i\in\mathcal{I}=\{1,\cdots,K\}}$. 包含 $|\mathcal{P}|$ 个正样本的表示为 $\mathcal{P}=\{p: p\in\mathcal{I}, y_p^s = y_i^s \wedge p\neq i\}$. 样本 i 与样本 p 的相关性计算如下:

$$\ell_{i,p} = \frac{\exp\left(\dfrac{\text{sim}(\boldsymbol{g}_i, \boldsymbol{g}_p)}{\tau}\right)}{\displaystyle\sum_{k\in\mathcal{I}\neq i}\exp\left(\dfrac{\text{sim}(\boldsymbol{g}_i, \boldsymbol{g}_k)}{\tau}\right)} \tag{3}$$

式中, $\boldsymbol{g}_i$ 为 i 个实例的整体语义表示; $\text{sim}(\boldsymbol{g}_i, \boldsymbol{g}_p) = (\boldsymbol{g}_i^{\mathrm{T}}\boldsymbol{g}_p)$ 为 2 个向量间的余弦距离; $\ell_{i,p}$ 为样本 i 与样本 p 的相似性; τ 为温度超参数, τ 值越大, 点积越

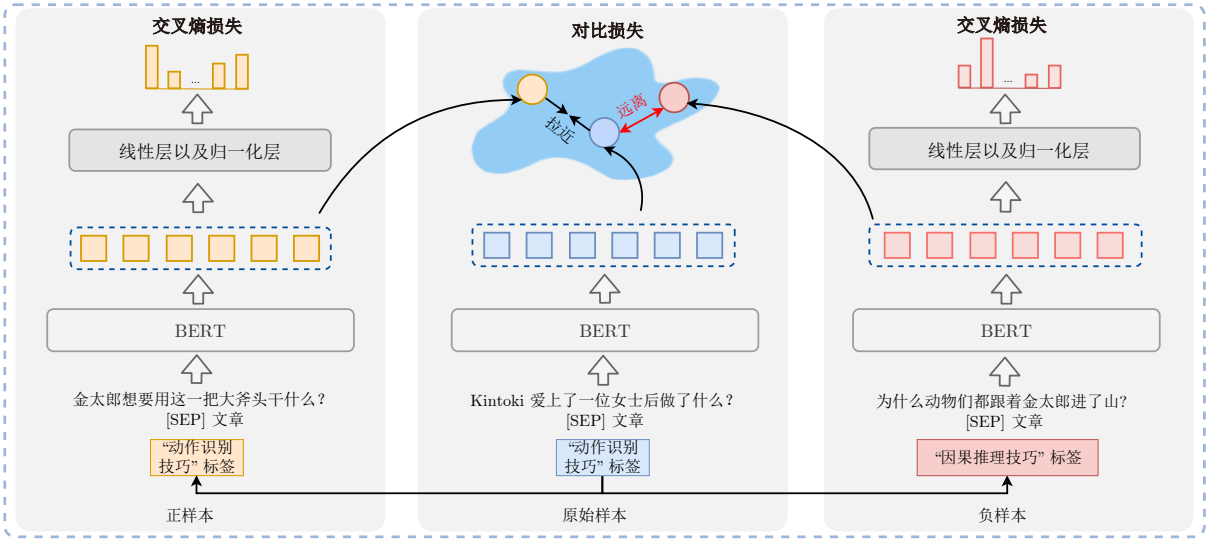


图 3　阅读技巧识别器结构

Fig. 3　Structure of the reading skill recognizer

小, 比较越困难.

在得到 $\ell_{i,p}$ 后, 对于批次 $\mathcal{I}$ 中的每个样本, 得到对比学习损失 $\mathcal{L}_{SCL}$:

$$\mathcal{L}_{SCL}=\sum_{i\in\mathcal{I}}-\ln\frac{1}{|\mathcal{P}|}\sum_{p\in\mathcal{P}}\ell_{i,p} \tag{4}$$

为了使 $\mathcal{L}_{SCL}$ 损失最小化, 应使正样本的相似度尽可能大, 不同类样本的相似度尽可能小. 所以阅读技巧识别器可在语义空间中将正样本聚在一起, 同时将不同类别的样本簇分开.

除了对比学习 (主要侧重于将样本与其他不同技巧样本进行聚合和分离), 本文还利用交叉熵损失来提高模型对各种类别样本的判别能力. 因此, 本文采用标准交叉熵损失 $\mathcal{L}_{CE}$ 进一步优化:

$$\mathcal{L}_{CE}=\text{CrossEntropy}\left(\text{MLP}(\boldsymbol{g}),\, y^s\right) \tag{5}$$

式中, CrossEntropy 表示交叉熵损失函数, MLP 表示多层线性感知机 (Multi-layer linear perceptron, MLP). 第 1 个阶段的总体损失 $\mathcal{L}_S$ 为 $\mathcal{L}_{SCL}$ 和 $\mathcal{L}_{CE}$ 的加权值:

$$\mathcal{L}_S=\mathcal{L}_{SCL}+\alpha\mathcal{L}_{CE} \tag{6}$$

式中, α 是用于平衡 2 个目标的超参数.

2.4 上下文编码器

第 2 个阶段进行机器阅读理解任务, 同时联合微调阅读技巧识别器.

上下文编码器利用预训练语言模型[34] (如 BERT) 来捕获问题和文章的上下文信息. 类似地, 上下文编码器会编码每个词, 以获得上下文信息向量 $(\boldsymbol{h}'_1, \boldsymbol{h}'_2, \cdots, \boldsymbol{h}'_T)$ (如式 (1) 所示). 为了使模型能够根据不同阅读技巧自适应地回答不同问题, 本文将技巧感知的语义表示 $\boldsymbol{h}_t$ 和上下文信息 $\boldsymbol{h}'_t$ 进行融合, 来对后续的答案预测进行指导.

2.5 双通道融合机制

双通道融合机制的目的是将上下文信息和技巧语义表示进行融合, 从而在辅助问答的过程中, 模型可以选择使用何种阅读技巧. 因此, 本文设计了拼接、多层线性感知机和协同注意力机制 3 种融合方法 (如图 4 所示), 并通过实验对比 3 种方法的性能.

2.5.1 拼接

拼接是一种非常简单并有效的融合 2 类信息的方法. 本文直接把阅读技巧的语义表示和上下文信息进行拼接, 得到更新后的表示为:

$$\boldsymbol{g}_t=\text{Concat}\left(\boldsymbol{h}_t,\, \boldsymbol{h}'_t\right) \tag{7}$$

2.5.2 多层线性感知机

多层线性感知机能通过神经网络有效地提取向量的抽象语义信息, 并能学习到 2 类信息间的相互依赖关系, 因此本文将 2 类信息通过多层线性感知机获得更新后的表示为:

$$\boldsymbol{g}_t=\text{MLP}([\boldsymbol{h}_t;\boldsymbol{h}'_t]) \tag{8}$$

2.5.3 协同注意力机制

协同注意力机制[2, 4] 被广泛应用在自然语言处理任务中, 它能有效地对不同信息源进行交互建模. 受注意力机制的启发, 本文利用协同注意力机制挖掘阅读技巧语义和上下文信息语义间的依赖关系. 在这一机制中, 注意力能够从 2 个方向计算: 一是从 $(\boldsymbol{h}_1, \boldsymbol{h}_2, \cdots, \boldsymbol{h}_T)$ 到 $(\boldsymbol{h}'_1, \boldsymbol{h}'_2, \cdots, \boldsymbol{h}'_T)$, 二是从 $\boldsymbol{h}'_t$ 到 $\boldsymbol{h}_t$:

$$\boldsymbol{g}_t=\text{attn}(\boldsymbol{h}_t(\boldsymbol{h}'_j)^{\text{T}})\boldsymbol{h}_t+\text{attn}(\boldsymbol{h}'_j(\boldsymbol{h}_t)^{\text{T}})\boldsymbol{h}'_j \tag{9}$$

式中, $\text{attn}(\cdot)$ 表示的是对 2 个向量相似度打分的计算函数, 本文采用余弦距离.

2.6 答案生成器

答案生成器的目的是根据双通道融合机制更新后的表示, 以自回归形式生成下一时刻词的概率分布. 对每个解码时刻 z, 答案生成器将已生成词的词向量 $\boldsymbol{W}_{y<z}$ 作为输入, 同时依赖双通道融合机制更

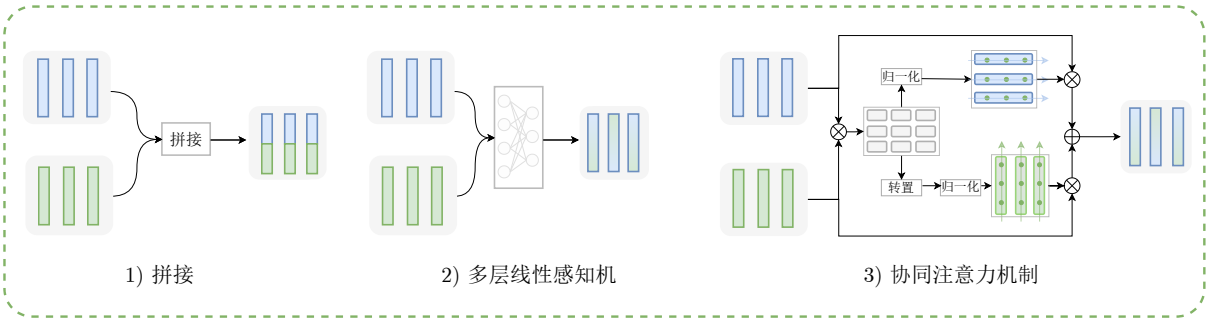


图 4 双通道融合机制结构图

Fig. 4 Structure of dual channel fusion machanism

新后的表示进行解码:

$$\boldsymbol{p}(y_z|\{y_1,\cdots,y_{z-1}\},\boldsymbol{g}_t)=\text{Decoder}(\boldsymbol{W}_{y<z},\boldsymbol{g}_t) \tag{10}$$

式中, Decoder 表示标准 Transformer 中的解码器. 本文使用双通道融合机制更新之后的语义向量 $\boldsymbol{g}_t$ 进行交叉注意力.

在第 2 阶段, 本文采用标准的交叉熵损失函数 $\mathcal{L}_{gen}$ 作为最终目标对模型进行优化:

$$\mathcal{L}_{gen}=-\sum_{z=1}^{Z}\ln \boldsymbol{p}(y_z|\{y_1,\cdots,y_{z-1}\},\boldsymbol{g}_t) \tag{11}$$

3　实验设置和结果

3.1　数据集

在 FairytaleQA 数据集上评估本文模型效果. 该数据集中含有明确的阅读技巧标签, 是一个专注于幼儿园到 8 年级学生叙事理解的数据集, 由教育专家人工标注. 该数据集由 278 个儿童故事中的 10580 个问题组成, 涵盖 7 种类型的叙事阅读技巧. 按照官方数据集的划分比例, 训练集、验证集、测试集以 8:1:1 的比例进行拆分. FairytaleQA 数据集的主要统计数据如表 1 所示.

表 1　FairytaleQA 数据集的主要统计数据

Table 1　Core statistics of the FairytaleQA dataset

项目	均值	标准偏差	最小值	最大值
每个故事章节数	15.6	9.8	2	60
每个故事单词数	2 305.4	1 480.8	228	7 577
每个章节单词数	147.7	60.0	12	447
每个故事问题数	41.7	29.1	5	161
每个章节问题数	2.9	2.4	0	18
每个问题单词数	10.5	3.2	3	27
每个答案单词数	7.2	5.8	1	70

3.2　阅读技巧描述

为了使模型能对不同问题采用不同的阅读技巧, 阅读技巧识别器通过有监督对比学习准确地捕捉它们的语义表示. 每种阅读技巧的定义如下: 1) 人物. 问题要求考生识别故事的人物性格或描述人物的特征属性. 2) 时间地点设置. 问题询问故事事件发生的地点或时间, 通常以“何时”或“何地”开头. 3) 动作. 问题询问角色的行为或有关该行为的信息. 4) 感受. 问题询问角色的情绪状态或对某些事件的反应, 通常措辞为“你感觉如何”. 5) 因果关系. 问题关注 2 个因果相关的事件, 其中前一事件导致问题中的后一事件. 这类问题通常以“为什么”开头. 6) 事件结果. 问题要求识别问题中先前事件导致的结果事件. 7) 事件预测. 问题询问事件的未知结果, 根据文本现有信息可以进行预测.

3.3　评价指标

1) 答案的评价指标包括 BLEU-n (B-n)[35]、ROUGE-L[36] 和 METEOR[37] , 它们被广泛用于评价自然语言生成任务中文本的质量; 2) 使用模型预测的准确率对阅读技巧识别进行评价. 指标的评价单位为 %.

3.4　实验设置

根据文献 [6], 相比其他模型, BART[38] 取得了最好结果, 因此本文直接将 BART 作为基线进行微调, 本文模型结构采取 Pytorch 框架实现. 模型的迭代轮数设置为 10 轮, 并以 5×10^{-5} 的初始学习率和 100 步的线性热身进行训练. 训练过程的批处理数据设置为 1. AdamW[39] 优化器的参数设置为 $\beta_1=0.900$、$\beta_2=0.999$ 和 $\epsilon=1\times10^{-8}$. 超参数设置为 $\alpha=0.5$. 对于第 1 阶段的预训练批处理数据大小设置为 32, 学习率设置为 5×10^{-5}, 温度超参数 $\tau=0.05$. 本文使用 Tesla V100-16 G GPU 进行实验.

3.5　基线模型

本文提供的基线包括轻量级模型 (用 FairytaleQA 数据集训练)、预训练语言模型 (直接评估, 按照文献 [6] 方法, 目的是进行比较) 和微调模型 3 种类型, 具体如下: 1) Seq2Seq[40] 是一个广泛用于自然语言生成任务的 RNN 编码器–解码器框架; 2) CAQA-LSTM[24] 是融入了问题类型信息的基于长短期记忆 (Long short term memory, LSTM) 网络的问答系统; 3) Transformer[41] 为标准的基于多头自注意力的 Transformer 模型; 4) BERT[34] 为预训练语言模型, 通过在所有层中联合调节上下文信息来学习深层次的双向表示; 5) DistilBERT[42] 为 BERT 的变体, 是一个较小的通用语言表示模型; 6) BART[38] 用于预训练序列到序列模型的去噪自动编码器, 在对文本生成进行微调时特别有效, 但也适用于理解任务; 7) BART-Question-types[22] 通过将问题类型及其类型定义结合到输入中来改进 FairytaleQA, 微调 BART 模型, 改进阅读理解系统的效果; 8) CAQA-BART[24] 融入了问题类型信息的基于 BART 的问答系统; 9) BART-NarrativeQA[6] 是在 NarrativeQA 数据集上进行微调过的变

体模型 BART, 是文献 [43] 中的一个基线模型; 10) BART-FairytaleQA[6] 是在 FairytaleQA 数据集上进行微调过的变体模型 BART, 是目前表现效果最好的模型. 值得注意的是, 本文设计了 2 种变体模型进行比较. 第 1 种方法是 BART-FairytaleQA †, 直接用 FairytaleQA 数据集进行微调; 第 2 种方法是在第 1 种方法基础上, 添加了技巧识别损失, 进行多任务联调, 用 BART-FairytaleQA ‡ 表示.

3.6 实验结果

FairytaleQA 数据集中验证集和测试集上的性能对比如表 2 所示, 评价的性能指标包括 Blue-n、ROUGE-L 和 METEOR. 其中 BART-FairytaleQA † 是文献 [6] 模型, 即最好的模型. BART-FairytaleQA ‡ 是在 BART-FairytaleQA † 基础上, 增加了阅读技巧识别损失.

由表 2 可以看出, 与其他模型相比, 本文模型在所有评估指标上都获得了最好的性能表现. 在测试集中, 本文模型在 Blue-4 和 ROUGE-L 指标上分别提高了 3% 和 4%. 值得注意的是, BART-FairytaleQA † 的 ROUGE-L 与人类表现结果的差距为 11.22%, 本文模型在 ROUGE-L 上的效果差距缩小为 6.11%, 取得了显著进步, 验证了本文模型的有效性. 实验结果表明, 阅读技巧识别是有利于模型回答更准确答案. 同时, 如何将阅读技巧的语义信息和上下文信息进行融合也是关键之一.

4 实验分析

4.1 消融分析

为了更好地了解模型中各组件的作用, 本文进行了消融实验, 实验结果如表 3 所示. 实验结果表明, 每个模块对最终结果都有提升效果. 表 3 中 SOTA (State of the art) 模型为最优模型.

4.1.1 去除阅读技巧识别器

阅读技巧识别器的目的是利用有监督对比学习损失, 准确地捕获阅读技巧的语义信息. 为了验证阅读技巧识别器的有效性, 本文去除阅读技巧识别器进行实验. 由表 3 可以看出, 去除阅读技巧识别器后, 所有指标都发生了显著下降, 验证了针对不同问题感知不同阅读技巧是非常有必要的.

4.1.2 去除对比学习损失

相比交叉熵损失建立在熵的基础上, 通常计算 2 个概率分布间的差值. 对比学习的目的是将属于同一类的样本的语义表示拉近, 并将不相关的样本分开. 为了验证模型的性能, 本文去除对比学习损失 (即去除式 (4)) 进行实验. 由表 3 可以看出, 模型的所有指标都有所下降, 验证了对比学习能更精确地识别不同的阅读技巧. 此外, 本文也验证了去除对比学习后, 相比仅基于交叉熵损失方法, 模型在 2 个任务上的结果都有所下降, 详见第 4.4 节.

表 2 FairytaleQA 数据集中验证集和测试集上的性能对比 (%)

Table 2 Performance comparison on the validation and the test set in FairytaleQA dataset (%)

模型名称	验证集						测试集					
	B-1	B-2	B-3	B-4	ROUGE-L	METEOR	B-1	B-2	B-3	B-4	ROUGE-L	METEOR
轻量化模型												
Seq2Seq	25.12	6.67	2.01	0.81	13.61	6.94	26.33	6.72	2.17	0.81	14.55	7.34
CAQA-LSTM	28.05	8.24	3.66	1.57	16.15	8.11	30.04	8.85	4.17	1.98	17.33	8.60
Transformer	21.87	4.94	1.53	0.59	10.32	6.01	21.72	5.21	1.74	0.67	10.27	6.22
预训练语言模型												
DistilBERT	—	—	—	—	9.70	—	—	—	—	—	8.20	—
BERT	—	—	—	—	10.40	—	—	—	—	—	9.70	—
BART	19.13	7.92	3.42	2.14	12.25	6.51	21.05	8.93	3.90	2.52	12.66	6.70
微调模型												
BART-Question-types	—	—	—	—	—	—	—	—	—	—	49.10	—
CAQA-BART	52.59	44.17	42.76	40.07	53.20	28.31	55.73	47.00	43.68	40.45	55.13	28.80
BART-NarrativeQA	45.34	39.17	36.33	34.10	47.39	24.65	48.13	41.50	38.26	36.97	49.16	26.93
BART-FairytaleQA †	51.74	43.30	41.23	38.29	53.88	27.09	54.04	45.98	42.08	39.46	53.64	27.45
BART-FairytaleQA ‡	51.28	43.96	41.51	39.05	54.11	26.86	54.82	46.37	43.02	39.71	54.44	27.82
本文模型	54.21	47.38	44.65	43.02	58.99	29.70	57.36	49.55	46.23	42.91	58.48	30.93
人类表现	—	—	—	—	65.10	—	—	—	—	—	64.40	—

表 3　FairytaleQA 数据集中验证集和测试集上的各组件消融实验结果 (%)
Table 3　The performance of ablation study on each component in our model on the validation set and the test set of the FairytaleQA dataset (%)

模型设置	验证集						测试集					
	B-1	B-2	B-3	B-4	ROUGE-L	METEOR	B-1	B-2	B-3	B-4	ROUGE-L	METEOR
SOTA 模型	51.28	43.96	41.51	39.05	54.11	26.86	54.82	46.37	43.02	39.71	54.44	27.82
去除阅读技巧识别器	52.15	44.47	42.11	40.73	55.38	27.45	54.90	47.16	43.55	40.67	56.48	29.31
去除对比学习损失	53.20	45.07	42.88	41.94	56.75	28.15	55.22	47.98	44.13	41.42	57.34	30.20
去除双通道融合机制	52.58	45.38	43.15	41.62	57.22	27.75	55.79	48.20	44.96	41.28	57.12	29.88
本文模型	54.21	47.38	44.65	43.02	58.99	29.70	57.36	49.55	46.23	42.91	58.48	30.93

4.1.3　去除双通道融合机制

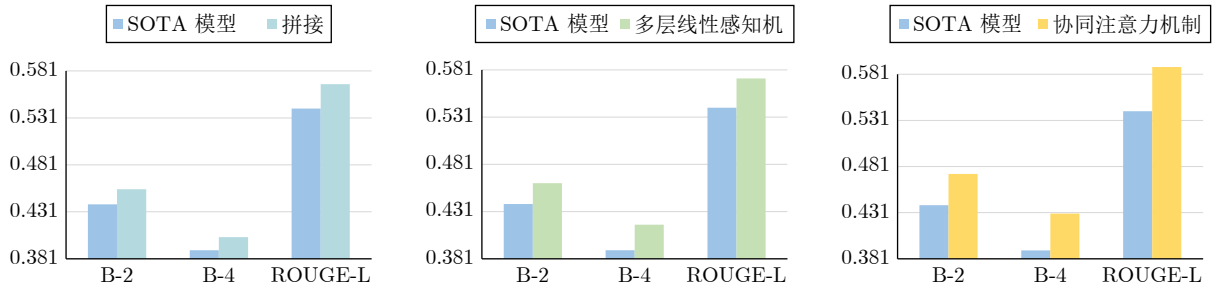
双通道融合机制的目的是更好地对上下文信息和阅读技巧的语义信息进行交互, 为了检验其对整个模型效果的表现, 本文去除双通道融合机制, 采用相加的方式来获得融合后的表示信息. 由表 3 可以看出, 去除双通道融合机制导致模型的所有指标都有所下降, 验证了该模块的有效性.

4.2　双通道融合机制分析

双通道融合机制的目标在于挖掘 2 种类型的交互信息, 本文中指上下文信息和阅读技巧的语义信息. 为了更进一步地探究拼接、多层线性感知机和协同注意力机制的有效性, 本文横向比较 3 种交互机制和基线 SOTA 模型的效果、纵向比较 3 种融合机制间的效果 2 个角度进行对比. 如图 5 所示, 验证了对上下文信息和阅读技巧的语义信息融合方法能带来一致性改进, 并在所有指标上均超过了基线 SOTA 模型. 图 5(a)、5(b)、5(c) 分别表示基线 SOTA 模型和拼接、基线 SOTA 模型和多层线性感知机、基线 SOTA 模型和协同注意力机制的实验结果, 其中协同注意力机制表现效果最好, 在 ROUGE-L 指标上增长了 5%. 原因是通过彼此的注意力能让模型感知到和阅读技巧更相关的上下文信息, 同时也可以利用上下文信息对阅读技巧的语义信息进行进一步增强和过滤. 图 6 展示了 3 种不同融合机制的比较. 由图 5 和图 6 可以看出, 在捕获到阅读技巧的语义信息后, 如何对 2 种信息进行融合, 仍值得进一步研究. 同时, 也验证了本文方法的有效性.

4.3　对比学习损失可视化分析

为了验证由于获得了更好的嵌入空间可以使阅读技巧识别性能提高, 对本文的有监督对比学习中嵌入的主成分分析 (Principal component analysis, PCA) 投影 $\boldsymbol{g}$ (见式 (2)), 并使用交叉熵损失的结果进行可视化操作作为反面对比, 本文还给出了原始数据空间的 PCA 投影 (见图 7(c)). 如图 7 所示, 使用有监督对比模型和有交叉熵损失模型都比原始数据更好地聚类具有相同标签的样本, 但基于有监督对比模型的嵌入空间中样本更易于区分, 不同类别的样本在语义空间中能够进一步分离. 此外, 基于交叉熵损失模型只聚类了 6 个样本簇, 而不是 7 个样本簇. 总之, 嵌入空间中样本的区分性越大, 阅读技巧识别和机器阅读理解的性能改善越好.



(a) SOTA 模型和拼接的实验结果
(a) Experimental results of SOTA model and concat operations

(b) SOTA 模型和 MLP 的实验结果
(b) Experimental results of SOTA model and MLP operations

(c) SOTA 模型和协同注意力机制的实验结果
(c) Experimental results of SOTA model and co-attention operations

图 5　双通道融合机制的性能比较
Fig. 5　The performances comparison on the dual channel fusion mechanism

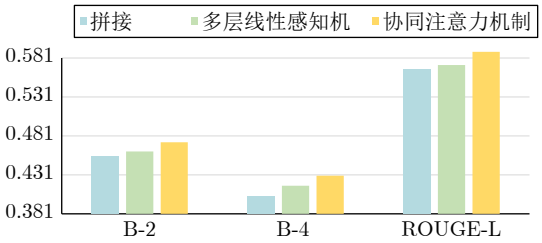


图 6 3 种不同融合机制的比较

Fig. 6 Comparison of the three different fusion mechanisms

4.4 阅读技巧识别分析

为了探究对比学习是如何影响阅读技巧识别性能, 从而影响机器阅读理解任务的, 表 4 中展示了基于交叉熵损失的方法和本文基于有监督对比学习损失的方法 (去除式 (4) 即去除 $\mathcal{L}_{SCL}$) 在 2 个任务上的效果. 可以看出, 本文模型在阅读技巧识别和机器阅读理解上都获得了令人满意的性能表现. 相比基于交叉熵损失的方法, 本文模型获得了更好结果, 在准确率指标上获得了 2.37% 的提高. 阅读技巧识别的较高准确性表明了有监督对比学习能够进一步帮助模型做出准确判断, 这也为机器阅读理解做出了指导. 另外, 通过第 4.3 节的对比学习损失可视化分析, 也能进一步验证引入对比学习的重要性.

4.5 阅读技巧识别器输入分析

为了研究阅读技巧识别器在不同输入上的表现, 在以下设置中进行分析: 1) 阅读技巧识别器只输入问题; 2) 阅读技巧识别器输入问题和文章. 实验结果见表 5, 可以看出, 当输入问题和文章时, 阅读技巧识别器的表现更好. 阅读技巧的学习不仅取决于问题, 还取决于问题与文章间的关系. 从人类表现角度看, 阅读技巧可以很容易地通过问题来预测 (因为人类已经学到了很多知识), 但对没有先验知识的机器来说, 问题通常很短, 导致信息较少. 通过引入文章内容, 信息得到了扩充, 使机器更容易学习上下文信息, 从而进行更准确的阅读技巧识别. 此外, 这种设置确保了 2 个训练阶段的输入是一致的, 可以缓解不同训练阶段造成的输入不同的差距.

4.6 样例分析

本文模型和基线模型的样例分析如下.

文章 1. 国王有个女儿, 和她死去的母亲一样漂亮, 还有一头金黄色的头发. 她长大了……她将成为王后, 因为国王的女儿和她死去的母亲长得一模一样. 我死后, 她的丈夫将……

问题: 国王的女儿长得怎么样?

预测阅读技巧: 人物感知技巧.

标签: 人物感知技巧.

答案: 和她死去的母亲一样漂亮.

模型预测:

1) 本文模型: 和她死去的母亲一样漂亮, 还有一头金黄色的头发.

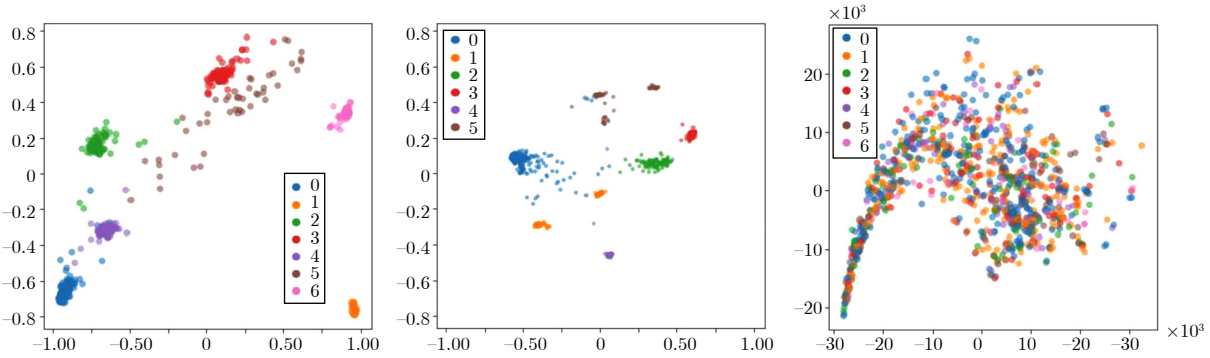
2) BART-FairytaleQA ‡: 和她死去的母亲长得一模一样.

3) Transformer: 长得一模一样.

4) Seq2Seq: 国王的女儿长得.

文章 2. 老人和妇人有两头光滑的母牛, 五只母鸡和一只公鸡, 一只老猫和两只小猫. 老人把时间都花在照看奶牛, 母鸡和花园上; 而妇人则忙着纺纱……

问题: 老人把时间花在了什么上?



(a) 使用有监督对比损失学习到的投影
(a) Projection learned using supervised contrastive loss

(b) 有交叉熵损失和没有对比损失的投影
(b) Projections with cross-entropy-loss and without contrastive loss

(c) 原始数据空间的 PCA 投影
(c) PCA projection of original data space

图 7 阅读技巧识别的可视化

Fig. 7 The visualization of the reading skill recognition

表 4　基于交叉熵损失的方法和基于有监督对比学习的方法在 2 个任务上的效果 (%)

Table 4　The performance of cross-entropy-loss-based method and supervised contrastive learning method on the two tasks (%)

实验设置	准确率	B-4	ROUGE-L	METEOR
基于交叉熵损失的方法	91.40	41.42	57.34	30.20
本文基于有监督对比学习损失的方法	**93.77**	**42.91**	**58.48**	**30.93**

表 5　不同输入下的阅读技巧识别器的识别准确率 (%)

Table 5　The recognition accuracy of reading skill recognizer under different inputs (%)

实验设置	验证集	测试集
只输入问题	85.31	82.56
输入问题和文章	**92.24**	**93.77**

预测阅读技巧: 动作识别技巧.

标签: 动作识别技巧.

答案: 老人把时间都花在照看奶牛、母鸡和花园上.

模型预测:

1) 本文模型: 老人把时间都花在照看奶牛、母鸡和花园上.

2) BART-FairytaleQA ‡: 老人和妇人有两头光滑的母牛, 五只母鸡和一只公鸡.

3) Transformer: 照看奶.

4) Seq2Seq: 老人把时间都花在看园上.

文章 1、文章 2 定性地显示了本文模型和基线模型的示例. 在文章 1 中, 本文模型预测了“人物”标签. 因此, 它生成的答案“和她死去的母亲一样漂亮, 还有一头金黄色的头发”是和类别标签“人物”相一致的, 而其他模型因为匹配到了问题中的“长得”关键词, 因此只匹配到了文章中的语句“和他死去的母亲长得一模一样”, 缺少了“漂亮”和“金黄色的头发”等描述. 然而, 本文模型描述了人物的属性信息, 并与真实的阅读技巧标签一致, 此外可以看出, 本文模型生成的答案更完整, 不仅包含了真实答案, 还补充了人物的特征信息“一头金黄色的头发”. 此外, 其他基线模型均产生了错误的答案和不完整的答案. 同样地, 在文章 2 中, 本文模型预测了“动作”标签, 并输出了符合“动作”阅读技巧标签的正确答案. 而其他模型由于无法准确捕获到阅读技巧信息, 因此生成的答案“……有两头光滑的母牛, ……”是和“动作”技巧不相关的. 总之, 各基线模型效果不佳. 样本分析实验证实挖掘不同的阅读技巧对于问答不同的问题至关重要, 能够辅助模型生成更加准确的答案.

5　结束语

本文以人类阅读过程为出发点, 对阅读理解任务中的问题进行更加细度分析, 认为针对不同的问题, 模型应该知晓采取何种策略来生成答案. 提出一种基于阅读技巧识别和双通道融合的机器阅读理解方法, 显式地捕获阅读技巧的语义信息, 并将其和上下文信息进行深层次融合. 实验结果表明, 本文方法实现了最先进的性能, 提高了技巧识别结果和机器阅读理解任务的性能. 样例分析验证了本文方法中各组件的重要性和所采用方法的有效性. 未来将研究如何挖掘问题背后更多的语义信息, 并提高阅读技巧识别的准确性能. 此外, 考虑到目前没有相关的数据具有标注信息, 而人工标注成本较大、时间开销较长, 未来会考虑在类似数据集上先进行人工标注, 再对模型进行实验. 也可通过聚类方法, 把拥有相同阅读技巧的文本进行聚类, 然后学习它们的上下文信息, 以减轻人工标注成本和计算开销.

References

1　Hermann K M, Kociský T, Grefenstette E, Espeholt L, Kay W, Suleyman M, et al. Teaching machines to read and comprehend. In: Proceedings of the Neural Information Processing Systems. Montreal, Canada: 2015. 1693–1701

2　Seo M J, Kembhavi A, Farhadi A, Hajishirzi H. Bidirectional attention flow for machine comprehension. arXiv preprint arXiv: 1611.01603, 2016.

3　Tay Y, Wang S, Luu A T, Fu J, Phan M C, Yuan X, et al. Simple and effective curriculum pointer-generator networks for reading comprehension over long narratives. In: Proceedings of the Conference of the Association for Computational Linguistics. Florence, Italy: 2019. 4922–4931

4　Peng W, Hu Y, Yu J, Xing L X, Xie Y Q. APER: Adaptive evidence-driven reasoning network for machine reading comprehension with unanswerable questions. *Knowledge-Based Systems*, 2021, **229**: Article No. 107364

5　Perevalov A, Both A, Diefenbach D, Ngomo A N. Can machine translation be a reasonable alternative for multilingual question answering systems over knowledge graphs? In: Proceedings of the ACM Web Conference. Lyon, France: 2022. 977–986

6　Xu Y, Wang D, Yu M, Ritchie D, Yao B, Wu T, et al. Fantastic questions and where to find them: FairytaleQA——An authentic dataset for narrative comprehension. In: Proceedings of the Conference of the Association for Computational Linguistics. Dublin, Ireland: 2022. 447–460

7　Liu S, Zhang X, Zhang S, Wang H, Zhang W. Neural machine reading comprehension: Methods and trends. arXiv preprint arXiv: 1907.01118, 2019.

8　Yan M, Xia J, Wu C, Bi B, Zhao Z, Zhang J, et al. A deep cascade model for multi-document reading comprehension. In: Proceedings of the Conference on Artificial Intelligence. Honolulu, USA: 2019. 7354–7361

9　Liao J, Zhao X, Li X, Tang J, Ge B. Contrastive heterogeneous graphs learning for multi-hop machine reading comprehension. *World Wide Web*, 2022, **25**(3): 1469−1487

10　Lehnert W G. Human and computational question answering. *Cognitive Science*, 1977, **1**(1): 47−73

11　Kim Y. Why the simple view of reading is not simplistic: Unpacking component skills of reading using a direct and indirect effect model of reading. *Scientific Studies of Reading*, 2017,

21(4): 310−333

12 Sugawara S, Yokono H, Aizawa A. Prerequisite skills for reading comprehension: Multi-perspective analysis of MCTest datasets and systems. In: Proceedings of the Conference on Artificial Intelligence. San Francisco, USA: 2017. 3089–3096

13 Weston J, Bordes A, Chopra S, Mikolov T. Towards AI-complete question answering: A set of prerequisite toy tasks. arXiv preprint arXiv: 1502.05698, 2015.

14 Purves A C, Söter A, Takala S, Vähäpassi A. Towards a domain-referenced system for classifying composition assignments. *Research in the Teaching of English*, 1984: 385−416

15 Vähäpassi A. On the specification of the domain of school writing. *Afinlan Vuosikirja*, 1981: 85−107

16 Chen D, Bolton J, Manning C D. A thorough examination of the CNN/daily mail reading comprehension task. arXiv preprint arXiv: 1606.02858, 2016.

17 Rajpurkar P, Zhang J, Lopyrev K, Liang P. Squad: 100,000 + questions for machine comprehension of text. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. Austin, USA: 2016. 2383–2392

18 Richardson M, Renshaw E. MCTest: A challenge dataset for the open-domain machine comprehension of text. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. Washington, USA: 2013. 193–203

19 Kocisk′y T, Schwarz J, Blunsom P, Dyer C, Hermann K M, Melis G. The narrativeqa reading comprehension challenge. *Transactions of the Association for Computational Linguistics*, 2018, **6**: 317−328

20 Yang B, Mitchell T M. Leveraging knowledge bases in LSTMs for improving machine reading. In: Proceedings of the Conference of the Association for Computational Linguistics. Vancouver, Canada: 2017. 1436–1446

21 Zhang Z, Wu Y, Zhou J, Duan S, Zhao H, Wang R. SG-Net: Syntax-guided machine reading comprehension. In: Proceedings of the Conference on Artificial Intelligence. New York, USA: 2020. 9636–9643

22 Kao K Y, Chang C H. Applying information extraction to storybook question and answer generation. In: Proceedings of the Conference on Computational Linguistics and Speech Processing. Taipei, China: 2022. 289–298

23 Lu J, Sun X, Li B, Bo L, Zhang T. BEAT: Considering question types for bug question answering via templates. *Knowledge-Based Systems*, 2021, **225**: Article No. 107098

24 Yang C, Jiang M, Jiang B, Zhou W, Li K. Co-attention network with question type for visual question answering. *IEEE Access*, 2019, (7): 40771−40781

25 Wu Z, Xiong Y, Yu S X, Lin D. Unsupervised feature learning via non-parametric instance discrimination. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE Computer Society, 2018. 3733–3742

26 Chen X, He K. Exploring simple siamese representation learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Virtual Event: IEEE, 2021. 15750–15758

27 Yang J, Duan J, Tran S, Xu Y, Chanda S, Li Q C, et al. Vision-language pre-training with triple contrastive learning. arXiv preprint arXiv: 2202.10401, 2022.

28 Chopra S, Hadsell R, LeCun Y. Learning a similarity metric discriminatively, with application to face verification. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. San Diego, USA: IEEE Computer Society, 2005. 539–546

29 Weinberger K Q, Saul L K. Distance metric learning for large margin nearest neighbor classification. *Journal of Machine Learning Research*, 2009, **10**(2): 207−244

30 Gao T, Yao X, Chen D. SimCSE: Simple contrastive learning of sentence embeddings. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. Virtual Event: 2021. 6894–6910

31 Giorgi J M, Nitski O, Wang B, Bader G D. DeCLUTR: Deep contrastive learning for unsupervised textual representations. In: Proceedings of the Conference of the Association for Computational Linguistics. Virtual Event: 2021. 879–895

32 Khosla P, Teterwak P, Wang C, Sarna A, Tian Y, Krishnan D, et al. Supervised contrastive learning. In: Proceedings of the Neural Information Processing Systems. Virtual Event: 2020. 18661–18673

33 Li S, Hu X, Lin L, Wen L. Pair-level supervised contrastive learning for natural language inference. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing. Virtual Event: IEEE, 2022. 8237–8241

34 Devlin J, Chang M, Lee K, Toutanova K. BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of North American Chapter of the Association for Computational Linguistics. Minneapolis, USA: 2019. 4171–4186

35 Papineni K, Roukos S, Ward T, Zhu W. BLEU: A method for automatic evaluation of machine translation. In: Proceedings of the Conference of the Association for Computational Linguistics. Philadelphia, USA: 2002. 311–318

36 Lin C Y. ROUGE: A package for automatic evaluation of summaries. *Text Summarization Branches Out*, 2004: 74−81

37 Banerjee S, Lavie A. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In: Proceedings of the Conference of the Association for Computational Linguistics. Ann Arbor, USA: 2005. 65–72

38 Lewis M, Liu Y, Goyal N, Ghazvininejad M, Mohamed A, Levy O, et al. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In: Proceedings of the Association for Computational Linguistics. Virtual Event: 2020. 7871–7880

39 Loshchilov I, Hutter F. Fixing weight decay regularization in adam. arXiv preprint arXiv: 1711.05101, 2017.

40 Cho K, Merrienboer B, Bengio Y, Gulcehre C, Bahdanau D, Bougares F, et al. Learning phrase representations using RNN encoder-decoder for statistical machine translation. In: Proceedings of the Empirical Methods in Natural Language Processing. Doha, Qatar: 2014. 1724–1734

41 Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A, et al. Attention is all you need. In: Proceedings of the Neural Information Processing Systems. Long Beach, USA: 2017. 5998–6008

42 Sanh V, Debut L, Chaumond J, Wolf T. DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter. arXiv preprint arXiv: 1910.01108, 2019.

43 Mou X, Yang C, Yu M, Yao B, Guo X, Potdar S. Narrative question answering with cutting-edge open-domain QA techniques: A comprehensive study. *Transactions of the Association for Computational Linguistics*, 2021, **9**: 1032−1046

彭 伟 中关村实验室助理研究员. 2023 年获得中国科学院信息工程研究所博士学位. 主要研究方向为对话生成, 网络空间安全.
E-mail: pengwei@iie.ac.cn
(**PENG Wei** Assistant professor at Zhongguancun Laboratory. He received his Ph.D. degree from Institute of Information Engineering, Chinese Academy of Sciences in 2023. His research interest covers dialog generation and cyber security.)

胡 玥 中国科学院信息工程研究所研究员. 主要研究方向为自然语言处理, 人工智能. 本文通信作者.
E-mail: huyue@iie.ac.cn
(**HU Yue** Professor at the Institute of Information Engineering, Chinese Academy of Sciences. Her research interest covers natural language processing and artificial intelligence. Corresponding author of this paper.)

李运鹏 中国科学院信息工程研究所博士研究生. 2019 年获得山东大学学士学位. 主要研究方向为自然语言处理. E-mail: liyunpeng@iie.ac.cn
(**LI Yun-Peng** Ph.D. candidate at the Institute of Information Engineering, Chinese Academy of Sciences. He received his bachelor degree from Shandong University in 2019. His main research interest is natural language processing.)

谢玉强 2023 年获得中国科学院大学博士学位. 主要研究方向为自然语言处理, 认知建模.
E-mail: yuqiang.xie@kunlun-inc.com
(**XIE Yu-Qiang** He received his Ph.D. degree from University of Chinese Academy of Sciences in 2023. His research interest covers natural language processing and cognitive modeling.)

牛晨旭 中国科学院信息工程研究所博士研究生. 2021 年获得西安电子科技大学学士学位. 主要研究方向为自然语言处理.
E-mail: niuchenxu@iie.ac.cn
(**NIU Chen-Xu** Ph.D. candidate at the Institute of Information Engineering, Chinese Academy of Sciences. She received her bachelor degree from Xidian University in 2021. Her main research interest is natural language processing.)